10.14 כריית נתונים

10.14.1 מטרה

כריית מידע משמשת לשיפור קבלת ההחלטות על ידי מציאת תבניות ותובנות שימושיות בנתונים.

10.14.2 תיאור

כריית מידע הוא תהליך אנליטי הבוחן כמויות גדולות של נתונים מנקודות מבט שונות ומסכם את הנתונים בצורה כזו שמתגלים דפוסים ויחסים שימושיים.

התוצאות של טכניקות כריית מידע הן בדרך כלל מודלים מתמטיים או נוסחאות המתארים דפוסים ויחסי בסיס. מודלים אלה ניתנים להצגה לצורך קבלת החלטות אנושיות באמצעות לוחות המחוונים והדיווחים החזותיים, או למערכות קבלת החלטות אוטומטיות באמצעות מערכות לניהול החוקים העסקיים או למימוש פנימי בתוך מסד נתונים.

כריית מידע יכולה להיות מנוצלת גם בחקירות בפיקוח או ללא פיקוח. בחקירה בפיקוח, משתמשים יכולים להציב שאלה ולצפות לתשובה שיכולה להוביל את קבלת ההחלטות שלהם. חקירה ללא פיקוח היא תרגיל גילוי דפוס טהור שבו דפוס מתגלה, ולאחר מכן משמש לקבלת החלטות עסקיות.

כריית מידע הוא מונח כללי המכסה טכניקות תיאוריות, אבחוניות וניבוייות:

- **תיאורי:** כגון קיבוץ אשכולות, קל יותר לראות את הדפוסים בקבוצת נתונים, כגון קווי דמיון בין לקוחות
- **אבחוני**: כגון עצי החלטה או פילוח יכול להראות מדוע דפוס קיים, כגון המאפיינים של הלקוחות הרווחיים ביותר לארגון
- **חזוי:** כגון רגרסיה או רשתות עצביות יכול להראות עד כמה סביר שמשהו עשוי להיות אמיתי בעתיד, כגון לחזות את ההסתברות כי תביעה מסוימת היא הונאה.

בכל המקרים חשוב לשקול את המטרה של תרגיל כריית המידע ולהיות מוכנים למאמץ ניכר בהבטחת הסוג הנכון, נפח ואיכות הנתונים שאיתם ניתן לעבוד.

10.14.3 אלמנטים

1. גילוי דרישות

המטרה וההיקף של כריית מידע נקבעים גם מבחינת דרישות תומכות החלטה לצורך קבלת עסקית חשובה, או במונחים של תחום פונקציונלי שבו נתונים רלוונטיים יוכנסו לגילוי תבניות ספציפיות לתחום העניין. אסטרטגיה זו של כרית מידע מלמעלה למטה לעומת אסטרטגיית כרייה מלמטה למעלה מאפשר ל BA לבחור את המערך הנכון של טכניקות כריית המידע

טכניקות קבלת החלטות פורמליות משמשות להגדרת הדרישות לתהליכי כריית מידע מלמעלה למטה. לתהליך גילוי מלמטה למעלה, מומלץ להשתמש בתובנה שהתגלתה על מודלים קיימים של החלטה, המאפשרים שימוש מהיר ופריסה של התובנה.

כריית המידע התהליכים הם פרודוקטיביים כאשר הם מנוהלים בסביבה זריזה -Agile. הם מסייעים בסבבים מהירים, אישור, ופריסה תוך מתן בקרות לפרויקט.

2. הכנת נתונים: מערך נתונים אנליטי

כלי כריית מידע עובדים על מערך נתונים אנליטי. הנתונים נוצרים בדרך כלל על ידי מיזוג רשומות מכמה טבלאות או מקורות לתוך מערך נתונים רחב אחד. קבוצות חוזרות מתמזגות בדרך כלל למספר קבוצות של שדות. הנתונים עשויים להיות מחולצים פיזית לקובץ בפועל או לקובץ וירטואלי שנותר במסד הנתונים או מחסן הנתונים, כך שניתן לנתח אותו.
מערכי הנתונים האנליטיים מחולקים לקבוצה המשמשת לניתוח וקבוצה עצמאית לחלוטין המשמשת לאישור שהמודל שפותח אכן נכון.
נפח הנתונים יכול להיות גדול מאוד, לפעמים וכתוצאה מכך נוצר לפעמים צורך לעבוד עם דגימות בלבד שיש לשמור בקובץ נפרד.

3. ניתוח נתונים

לאחר שהנתונים זמינים, הם מנותחים. מפעילים מגוון רחב של אמצעים סטטיסטיים כולל כלים להדמיה המשמשים כדי לראות כיצד הנתונים מפוזרים, איזה נתונים חסרים, וכיצד מאפיינים מחושבים שונים מתנהגים. צעד זה הוא לעתים קרובות הארוך ביותר והמורכב ביותר במאמץ כריית המידע ולכן נעשים מאמצים לבצע אותו באמצעות מיקוד של אוטומציה. רוב כוחו של מאמץ כריית המידע מגיע בדרך כלל מזיהוי מאפיינים שימושיים בנתונים. לדוגמה, מאפיין עשוי להיות מספר הפעמים שהלקוח ביקר בחנות ב 80 הימים האחרונים. הקביעה כי ספירה במהלך 80 הימים האחרונים הוא יותר שימושי מאשר לספור 70 או 90 הוא המפתח להצלחה.

4. טכניקות מידול

יש מגוון רחב של טכניקות כריית מידע.

כמה דוגמאות של טכניקות כריית נתונים הן:
- סיווג ועוצמת רגרסיה (CART)
- טכניקות שונות לעצי החלטה כולל C5
- רשתות עצביות,
- רגרסיה ליניארית ולוגיסטית,
- predictive scorecards

הנתונים האנליטיים והמאפיינים המחושבים מוזנים לתוך אלגוריתמים אלו אשר אינם מפוקחים (המשתמש אינו יודע מה הוא מחפש) או בפיקוח (המשתמש מנסה למצוא או לחזות משהו ספציפי). טכניקות מרובות משמשות לעתים קרובות כדי לראות איזו היא היעילה ביותר. נתונים מסוימים מוחזקים מהמודלים ומשמשים לאישור שהתוצאה יכולה להיות מאומתת עם נתונים שלא נעשה בהם שימוש ביצירה הראשונית.

5. מימוש

לאחר בניית המודל, יש לממש אותו כדי שיהיה שימושי. מודלים של כריית מידע ניתנים למימוש במגוון דרכים, או כדי לתמוך במקבלי ההחלטות האנושיים או כדי לתמוך במערכות קבלת החלטות אוטומטיות. עבור משתמשים אנושיים, תוצאות כריית נתונים יכולות להיות מוצגות באמצעות מטאפורות חזותיות או כשדות נתונים פשוטים. טכניקות רבות של כריית מידע מזהות כללים עסקיים פוטנציאליים שניתן לממש באמצעות מערכת ניהול החוקים העסקיים. כמה טכניקות לכריית צידע - במיוחד אלו המתוארות כטכניקות אנליטיות מנבאות - יוצרות נוסחאות מתמטיות. אלה יכולים להיות ממומשות על ידי חוקים עסקיים אבל יכולים לשמש גם כדי ליצור SQL או קוד למימוש. מגוון רחב יותר של אפשרויות מימוש בתוך מסד נתונים מאפשר לשלב מודלים אלה בתשתית הנתונים של הארגון

10.14.4 שיקולי שימוש

1. חוזקות

• חושף דפוסים נסתרים ויוצר תובנות שימושיות במהלך הניתוח - מסייע לקבוע אילו נתונים כדאי לשמור או כמה אנשים עשויים להיות מושפעים מהצעות ספציפיות.

• ניתן לשלב בתכנון המערכת כדי להגביר את דיוק הנתונים.

• ניתן להשתמש בכרית מידע כדי למנוע או להפחית את ההטיה האנושית על ידי שימוש בנתונים כדי לקבוע את העובדות לאשורן.

2. מגבלות

• הפעלת כמה טכניקות ללא הבנה מדוייקת כיצד הם עובדים יכולים לגרום לתובנות מוטעות.

• גישה לנתונים גדולים ולכלי נתונים מתוחכמים לכריית מידע עלולים להוביל לשימוש לא נכון בשוגג.

• טכניקות וכלים רבים דורשים ידע של מומחה על מנת לעבוד איתם
• טכניקות מסוימות משתמשות במתמטיקה מתקדמת ברקע ובחלקן
בעלי העניין אינם יכולים לקבל תובנות ישירות לתוצאות. חוסר שקיפות נתפס עלול לגרום להתנגדות מצד חלק מבעלי העניין.\
• תוצאות כריית מידע עשויות להיות קשות למימוש, אם קבלת ההחלטות שהן נועדו להשפיע עליהן, אינה מובנת.

.